DOCUMENT RESUME

ED 248 270 TM 840 576

AUTHOR Busch, John Christian; Jaeger, Richard M.
TITLE An Evaluation of Methods for Setting Standards on the Essay Portion of the National Teacher Examinations.
PUB DATE Apr 84
NOTE 20p.; Paper presented at the Joint Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education (New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Entrance Examinations; Criterion Referenced Tests; *Cutting Scores; *Essay Tests; Evaluation Methods; Higher Education; Holistic Evaluation; *Interrater Reliability; Preservice Teacher Education; Scaling; *Scoring; Writing Evaluation
IDENTIFIERS Angoff Methods; Educational Testing Service; *National Teacher Examinations; North Carolina

ABSTRACT

This study addressed seven questions regarding the methods used in setting passing scores on the essay subtest of the National Teacher Examinations (NTE) Communication Skills test for the North Carolina State Board of Education. North Carolina uses these tests to screen prospective applicants to teacher education programs. The judges (five college faculty members and two public school teachers) rated two sets of eleven essay responses to a question from the 1982 national NTE administration provided by Educational Testing Service (ETS). The first essay set was presented in order of increasing ETS assigned scores, the second in random order. Judges rated the essays as "acceptable" or "unacceptable." They were given the ETS holistic scoring for each essay, decided which ETS rating represented a minimally acceptable essay, and rank ordered the essays. A minimal standard for each judge was computed by two methods. The resulting data were analyzed descriptively due to the small sample size. Results suggested that to produce the smallest variance among judges' standards, essays should be presented in increasing order of quality (as scored by ETS readers), but without prior knowledge of ETS ratings. Rating knowledge created an apparent positive bias in recommended standards. (BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED248270

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

AN EVALUATION OF METHODS FOR SETTING STANDARDS ON THE ESSAY PORTION OF THE NATIONAL TEACHER EXAMINATIONS

John Christian Busch and Richard M. Jaeger
University of North Carolina at Greensboro

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. C. Busch

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 840 526

Presented at the 1984 Annual Meeting of the American Educational Research Asscciation and the National Council on Measurement in Education, New Orleans, La. April 23-27, 1984, in a symposium entitled "Setting Standards for Teacher Certification: Recent Advances in Methodology."

BACKGROUND AND RATIONALE

Stalnaker (1951) has defined the essay question as, "...a test item which requires a response composed by the examinee, usually in the form of one or more sentences, of a nature that no single response or pattern of responses can be listed as correct, and the accuracy and quality of which can be judged subjectively only by one skilled or informed in the subject...not only can no single answer be listed as correct and complete...but even an expert cannot usually classify a response as categorically right or wrong"(p.495).

Much has been written about the advantages and weaknesses of, construction of, and scoring of essay tests (e.g., Coffman 1971, 271-302). One must assume that, despite difficulties in obtaining reliable evaluations of essays, when one wishes to evaluate writing skills,a direct measure of writing is essential and should be included in any battery which tests basic communication skills.

A recent issue of Educational Measurement: Issues and Practice is devoted to the assessment of writing. In that issue, Meredith and Williams(1984) cite a National Council of Teachers of English statement which argues that the most appropriate method of measuring writing skill is through direct assessment. This position is reflected in state policies in South Carolina, Maryland, Texas, Louisiana, New Jersey, Connecticut, Illinois, and North Carolina, each of which addresses writing assessment in recent legislation (Meredith and Williams, 1984).

Mullins (1984) identifies three methods used in scoring and evaluation of essays: holistic scoring, primary trait scoring and analytic scoring. Holistic scoring is the method used by Educational Testing Service (ETS) to score the essay subtest of the National Teacher Examinations (NTE). This method assigns scores on a six-point scale, following careful training and calibration of readers and systematic attempts to achieve consensus. Mullins points out that "For decisions about proficiency or competency, standards must be defined and additional analyses conducted to determine which score points meet various levels of quality (p. 17)."

In 1982 the North Carolina State Board of Education contracted with the Center for Educational Research and Evaluation at the University of North Carolina at Greensboro to plan and conduct a validation and standard-setting study of the National Teacher Examinations Communication Skills and General Knowledge Tests. North Carolina's intent in using these tests is to screen prospective applicants to teacher education programs in basic communication skills and general knowledge.

In seeking to provide a recommended standard for the Communication Skills test, it was first necessary to determine raw score standards for each of the four individual subtests comprising the Communication Skills test (multiple-choice tests in Listening, Reading, and Writing and an essay Writing test). These four scores are combined and then transformed to a standard-score scale using an ETS-provided equation.

Our initial research design (Jaeger and Busch, 1982) proposed a modification of the Angoff (1971) standard-setting procedure

for the multiple-choice subtests of the Communication Skills Test. This standard-setting procedure required item-focused probability judgments but was iterative, and provided judges with information to help them evaluate the consequences of their decisions. It was obvious that a different approach to standard setting would have to be used for the essay test. Angoff's procedure could not be used since judges could not make decisions about examinees' likely performance on individual test items.

Research Questions

This study addressed several questions regarding the methods to be used in setting passing scores on the essay subtest of the National Teacher Examinations Communication Skills Test.

1) What standards are recommended for the essay sub-test when judges evaluate a set of essays not knowing the ratings provided by ETS; i.e. they provide "uninformed" reviews?

2) What standards are recommended for the essay sub-test when the judges are informed of the ETS ratings ?

3) What is the effect on recommended essay sub-test standards (central tendency and variability) when the judges are "informed" compared to when they are "uninformed" of the ETS-assigned ratings?

4) What is the effect on essay sub-test standards (central tendency and variability) when a minimal passing score is computed by two different methods? (one which recognizes inversions in ratings versus one which does not)?

5) Are judges consistent in their evaluation of two inde-

pendent sets of essays?

6) How do rankings of essays provided by this study's judges compare to the ratings provided by Educational Testing Service?

7) What is the effect on recommended standards of systematic vs. random order of presentation of essays?

METHODS

Judges:

Judges for the study consisted of 5 college and university faculty members and 2 public school teachers. College and university faculty were randomly sampled from selected departments in five public and private colleges in or near Greensboro, North Carolina. Faculty assigned to review the Essay subtest were sampled from faculty teaching composition courses in departments of English or faculty who had assignments to teach writing as part of a special services (remedial) staff. Public school teachers who participated were sampled from the three school systems in Guilford County, North Carolina. To qualify for selection, teachers had to hold a current teaching certificate, be employed full time, be a graduate of a North Carolina teacher education program, and have between two and five years of teaching experience.

Materials

The judges rated two sets of essays that had been provided by Educational Testing Service. All essays were written in response

to the same question during the fall 1982 national administration of the NTE. These essays, eleven in each set, were selected to represent performance at different levels of quality. ETS scores its essays holistically, on a "1 to 6" scale with "1" the lowest score and "6" the highest. ETS has two graders score each essay, and the resulting scores are summed to provide essay scores that range from 2 (low) to 12 (high). The essays used in this study, represented a combination of scores at each of the eleven levels described in Figure 1. In the analyses reported here, the average of the ETS-assigned ratings was used, so as to produce scores on a scale from one to six.

INSERT FIGURE ONE ABOUT HERE

Procedures

Uninformed Judgments. Each of the judges was given the first set of eleven essays that had been provided by ETS. As each essay was read, the judge was asked to respond to the following directions:

> As you read each essay, I would like you to make a rather simple judgment: First, please assume that each essay has been written by someone who has just completed two years of typical course work at a university or college and who is applying for admission to a teacher education program.
>
> Second, answer the following question, "Given the standards which might be used for sophomore level students is this essay at least minimally acceptable for admission to teacher education." If your answer is yes, please mark "A" on your answer sheet. If your answer is no, the essay is not at least minimally acceptable then please mark choice "U" on your answer sheet. I should stress that the essays you identify as acceptable will include those essays which are just minimally acceptable. Please grade the essays in the order in which they are arranged.

As the research design was implemented, the second set of essays was presented in random order, and the first set was presented in order of increasing ETS-assigned scores.

Classifications of the essays as "acceptable" or "unacceptable" were used to compute a minimal standard for each judge. This was done in two ways:

Method A: Take as the minimal acceptable standard, the lowest ETS-assigned score on an essay classified as "acceptable.

Method B: This method is somewhat more complicated and uses an approach suggested by Lawrence Cross of Virginia Polytechnic Institute and State University. The method attempts to account for the distinct possiblity of "inversi " in the essays judged to be acceptable and unacceptable after they have been arranged in increasing order of ETS-assigned scores. An example of a set of essays in which there is such an inversion is given on the left of Figure 2. In this example, the essay rated 3.0 received an "unacceptable rating". If the essays are re-arranged in descending order of average ETS-assigned scores, but within judgment categories, i.e., with all "acceptable" essays listed above the solid line, and "unacceptable" essays listed below, the inversion becomes apparent. The upper dotted line separates essays that received an acceptable rating and whose average ETS-assigned scores are in an uninterrupted sequence (4.5, 4.0, 3.5,) from an "acceptable" essay that received an average ETS-assigned score of 2.5. The inversion occurs because the essay with an average ETS-assigned score of 3.0 was judged to be "unacceptable" by this judge, while the essay with an average ETS-assigned score of 2.5 was judged to be "acceptable". The lower dotted line separates the essays judged to be "unacceptable" into a lower group (2.0, 1.5) with average ETS-

assigned scores that are below each of the average ETS-assigned scores for essays judged to be acceptable, and the essay with an average ETS-assigned score of 3.0, that does not satisfy this criterion.

The difficulty in this situation is determining how one should best identify a minimally acceptable score on the ETS scale. One can use the average score associated with the lowest "acceptable" essay which is 2.5 (Method A). However, one can also focus on the highest average ETS-assigned score of the essays rated unacceptable (a score of 3.0). Since 3.0 is the highest average score assigned to unacceptable essays, one could assume that 3.5, the next highest scale point would be minimally acceptable. As a compromise, one could use as a standard, the average of these scale values (2.5 and 3.5) that is, 3.0. It is apparent that, in cases where there are inversions, Method B provides standards that are systematically higher than those resulting from Method A.

INSERT FIGURE 2 ABOUT HERE

Informed Judgments. The process of obtaining individual judgments used above was described as "uninformed" since the judges were not aware of the scores assigned to the essays by ETS readers. Immediately following the steps described above, the judges were informed of the ETS ratings for each essay. They were told:

The sheet being handed out lists the essay letter and the rating assigned by ETS to each. You'll note that each essay has two grades because it was evaluated by two evaluators. The

"holistic" grades assigned by ETS range on a scale that goes from 1 to 6, with 6 the most acceptable and 1 the least acceptable. It will probably help you to arrange the essays in order of increasing quality, as judged by ETS.

You will notice that there are essays at different points along the "1" to "6" scale The lowest-graded essay received a 1 from both graders for a total score of 2; the highest-graded essay received a score of 6 from both graders for a total score of 12. We recognize that your ratings might not be the same as those of the ETS raters. If you let the essays define different scores on the 1 to 6 scale please tell us, on that 1-to-6 scale used by ETS, the scale value which you would consider to represent the minimally acceptable essay for a sophomore applying for admission to teacher education. Please circle...the point on the 1-to-6 scale to indicate the point at which a minimally acceptable essay would fall.

When the judges completed this task, they were asked whether they agreed with the ETS rating of the essays, and if they did not, they were asked to rank the essays in order from highest to lowest quality.

Educational Testing Service provided two separate sets of essays and the process was repeated a second time. The essays in the second set were presented to judges in random order.

Statistical Analysis

The seven research questions listed above motivated several analyses of the data. The results of these analyses must be interpreted cautiously, since only seven judges participated in this review of the Essay portion of the Writing Test. Because of the obvious lack of power which would be present in an inferential analysis, our data analyses were totally descriptive.

RESULTS

The first research question concerned the standards that judges would recommend when they had not been informed of the ratings ETS readers had applied to the essays. Data pertinent to this question are shown in Table 1; standards for each of the seven judges are reported for both sets of essays. The recommended standards shown in this table have been computed using both Method A and Method B. The ETS one-to-six scale of standards has been used in computing these standards. Since the essays provided by ETS were scored by two readers, the scores assigned each essay have been averaged across readers.

INSERT TABLE 1 ABOUT HERE

In cases where judges were totally consistent in their ratings of essays (i.e. an essay given a lower score by an ETS reader was never judged to be "acceptable" at the same time an essay given a higher score by an ETS reader was judged to be "unacceptable"), the two computation methods resulted in identical standards. This was always the case for the first set of essays, which were presented to judges in order of increasing quality based on the average ETS-assigned scores. In cases where there were inversions, i.e., a judge rated an essay with a higher average ETS-assigned score as unacceptable and at the same time rated an essay with a lower score as acceptable, Method B resulted in higher recommended standards than did Method A. This result was observed for three of the seven judges when they rated the second set of essays. It should be noted that the second set of essays was presented to the judges in random order of average

scores assigned by ETS readers, rather than in increasing order of these scores.

Our fourth research question concerned the difference, if any, between standards resulting from use of the two methods of computing minimal standards. Table 2 contains data that bear on this question. In terms of average recommended standards, there is very little difference between the results computed using Methods A and B. For the first set of essays, as we have noted, the two methods produced identical results since there were no inversions. For the second set of essays, the average recommended standards differed by four-tenths of a point on the six-point ETS scale. When averaged across both sets of essays, the mean recommended standards resulting from the two methods differed by two-tenths of a point. Where there were differences, Method B resulted in a higher recommended standard than did Method A.

INSERT TABLE 2 ABOUT HERE

It is noteworthy that for Essay Set #2, Method B produced a smaller standard deviation of recommended standards (across judges) than did Method A. This positive feature is to be expected, since the standard produced by Method B is an average that can take on whole-number, as well as a range of fractional values, whereas the standard produced by Method A can only take on the finite set of scale values assigned to specific essays (in this application, the values 1.0, 1.5, 2.0, etc. up to 6.0).

The fifth research question asked whether judges would be

consistent in their recommended standards across the two sets of essays, in the sense that their recommended standards would be at similar scale points on the one-to-six scale. The data in Table 2 suggest that, on average, judges were quite consistent across the two sets of essays. When Method A was used, the average standards recommended for the two sets of essays differed by less than 0.3 scale points. When computational Method B was used, the average standard recommended for the two sets of essays differed by less than 0.2 scale points. These differences reflect not only the consistency of the judges used in this study, but the consistency of the scores assigned to the two sets of essays by ETS readers. They also show that the effects of the orders of presentation used in this study for the two sets of essays was small. In light of these components of variation, it appears that the standards were consistent.

Our second research question concerned the standards judges would recommend when they were informed of the scores the ETS graders had given each of the essays; the third question concers a comparison between the standards judges recommended when they were uninformed about the scores applied by ETS readers and when they were informed of these scores. Data pertinent to these questions are also provided in Table 2.

The average standards recommended by the seven judges when they were informed of the scores assigned by ETS readers were virtually identical for the two sets of essays; about 3.1 on the six-point scale. This average is about 0.6 points higher than the average recommended standard in the absence of information on the scores applied by ETS readers. It appears that judges are a

bit more lenient in their standards when they are not informed of scores assigned to the essays by other raters. The data in Table 2 also suggest that the variability of recommended standards is larger when judges are informed of scores assigned the essays than when they are not so informed. If these small-sample results can be generalized, it appears that somewhat more-reliable standards result when judges are not informed of the scores ETS readers have assigned to the essays.

There were very few "inversions" of average scores assigned to the essays by ETS readers. The sixth research question we posed, "How do rankings of essays provided by judges compare to the ratings provided by Educational Testing Service?" must be answered positively; i.e., the rankings provided by our judges and the rankings of average ETS-assigned scores were very similar.

Our seventh question asked whether any conclusions could be drawn about the ordering of the essays as they were presented to the judges. That is, should essays be randomly arranged or systematically arranged in terms of average ETS-assigned scores. An unplanned manipulation of the order of the essays resulted in one of the essay sets being systematically ordered and the other being randomly ordered. While this event produced a design which falls far from an optimal and planned experiment designed to test this question, we did inspect the standards to determine if any useful interpretations could be made. Acknowledging that there is a confounding of order with essay set, we note that, while the difference in central tendency of recommended standards between

the orders (essay sets) is minimal (about .10 scale points averaged across the two methods), the standard deviation of the recommended standards tends to be slightly smaller for the essay set presented in systematic order.

CONCLUSIONS

In summary, this small-sample study suggests that essays should be presented to judges in increasing order of quality, as scored by ETS readers, but without prior knowledge of the scores applied by ETS readers. This combination of factors appears to produce the smallest variance among the standards recommended by different judges, thus maximizing the reliability of the average recommended standard. This conclusion is also based on the apparent positive bias in recommended standards resulting from knowledge of the essay grades applied by ETS readers.

In addition, it should be remembered that the appropriate judgment task should be based on thoughtful consideration of the quality of the essays, not solely on consideration of the scores applied to those essays by other readers.

REFERENCES

Angoff, W. Scales, norms, and equivalent scores. In R.L.Thorndike (ed.) Educational Measurement. (2nd. ed.) Washington, D.C.: American Council on Education, 1971, 514-515.

Coffman, W. Essay examinations. In R.L.Thorndike (ed.) Educational Measurement. (2nd. ed.) Washington, D.C.: American Council on Education, 1971, 271-302.

Jaeger, R. & Busch, J. A research-based operational plan for validating and setting standards on the General Knowledge and Communication Skills subtests of the National Teacher Examinations. University of North Carolina at Greensboro: Center for Educational Research and Evaluation, 1982.

Meredith, V. & Williams, P. Issues in direct writing assessment: Problem identification and control. Educational Measurement: Issues and Practice, 1984, 3, 11-15.

Mullins, I. Scoring direct writing assessments: What are the alternatives? Educational Measurement: Issues and Practice, 1984, 3, 16-18.

Stalnaker, J. M. The essay type of examination. In E.F. Lindquist (ed.) Educational Measurement. (2nd. ed.) Washington,D.C.:

Figure 1. Range of Scores on Essays Provided by Educational Testing Service from the Fall 1982 National Administration of the National Teacher Examinations

Essay ID	ETS-Assigned Score#1	ETS-Assigned Score#2	Total ETS Score	Average of ETS-Assigned Score
1	6	6	12	6.0
2	6	5	11	5.5
3	5	5	10	5.0
4	5	4	9	4.5
5	4	4	8	4.0
6	3	4	7	3.5
7	3	3	6	3.0
8	3	2	5	2.5
9	2	2	4	2.0
10	2	1	3	1.5
11	1	1	2	1.0

Figure 2. Example of a Set of Essays With an Inversion in Ratings

Arranged in Decreasing Order of Averaged ETS-Assigned Score

ETS Values	Judgments
4.0	Acceptable
3.5	Acceptable
3.0	Unacceptable
2.5	Acceptable
2.0	Unacceptable
1.5	Unacceptable

Arranged in Decreasing Order of Average ETS-Assigned Scores Within Judgment Category

ETS Values	Judgments
4.0	Acceptable
3.5	Acceptable
2.5	Acceptable
3.0	Unacceptable
2.0	Unacceptable
1.5	Unacceptable

Table 1. Minimally Acceptable Scores Recommended by Seven Judges When Essay Standards are Computed by Two Methods

Essay Set #1

Method	Judge No:	1	2	3	4	5	6	7
Method A		2.0	2.5	2.5	3.0	2.5	2.5	3.0
Method B		2.0	2.5	2.5	3.0	2.5	2.5	3.0

Essay Set #2

Method	Judge No:	1	2	3	4	5	6	7
Method A		3.0	1.5	3.0	1.5	1.5	3.0	2.5
Method B		3.0	3.1	3.0	2.0	2.2	3.0	2.5

Table 2. Summary Statistics for Essay Standard-Setting Methods

Uninformed Judgment Methods

	1st Essay Set		2nd Essay Set		Average for Two Essay Sets	
	Method A	Method B	Method A	Method B	Method A	Method B
$\bar{X}$	2.57	2.57	2.29	2.69	2.43	2.63
s	.35	.35	.76	.45	.35	.18
min	2.00	2.00	1.50	2.00	2.00	2.35
max	3.00	3.00	3.00	3.10	2.75	2.80

Informed Judgment Methods

	1st Essay Set	2nd Essay Set	Average for Two Essay Sets
$\bar{X}$	3.14	3.07	3.12
s	.38	.61	.45
min	3.00	2.00	2.50
max	4.00	4.00	4.00